IT-assisted Exploration of Excavation Reports. Using Natural Language Processing in the archaeological research process

Chiarcos, C.; Lang, M; Verhagen, J.W.H.P.

published in
CAA 2015. Keep the Revolution Going. Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology
2016

document version
Publisher's PDF, also known as Version of record

IT-assisted Exploration of Excavation Reports. Using Natural Language Processing in the Archaeological Research Process

Christian Chiarcos
chiarcos@informatik.uni-frankfurt.de
Applied Computational Linguistics, Goethe-University Frankfurt am Main, Germany

Matthias Lang
matthias.lang@uni-tuebingen.de
eScience-Center, University of Tübingen

Philip Verhagen
j.w.h.p.verhagen@vu.nl
VU University Amsterdam

Abstract: In this paper we summarize recent experiments conducted on what has become known as 'Machine Reading' of scientific literature from different fields of archaeology, i.e., the extraction of machine readable, semantic information out of plain text. We describe a processing pipeline to extract semantic concepts and relations, the representation of extracted information by means of Semantic Web standards, its linking with background knowledge from both domain vocabularies and general lexical/conceptual knowledge sources and possible user interfaces that provide access to the extracted information.

These experiments represent early steps in the development of an elaborate system that will allow to analyse excavation reports, access/search them on a semantic, rather than a textual basis, and augmenting them with background information specific to the field of archaeology.

Keywords: Natural Language Processing, Semantic Web, Linked Open Data (LOD), Archaeological grey literature

1 Motivation

Archaeology as a scientific discipline was established during the 19th century, and with almost two centuries of scientific publications, one of the most time-consuming challenges to nearly every researcher nowadays consists in the task to assess this huge amount of knowledge created by earlier generations of scholars on a particular phenomenon, region or artefact type currently under consideration. Over time, many of them worked in different countries, with different methodological and ideological backgrounds, using different terminologies in different languages. Accordingly, an exhaustive overview over, say, the distribution of Roman coins in Celtic contexts, requires not only to cover an enormous wealth of literature, but also, a wealth of literature of extreme heterogeneity. In addition, great parts of this information may be available only as 'grey literature', as technical reports, in-house publications of different universities, thesis papers or mere manuscripts. In the digital age, this body of sparsely accessible knowledge grows even more drastically than the number of traditional print publications.

Machine reading represents itself as a way to improve the accessibility of this wealth (or, in parts, this mess) of information: the extraction of machine-readable, formalized knowledge out of written text, and ways to make this information accessible to scholars in the field in a way that it can be used without or with minimal technical expertise.

We describe one selected case study in this regard, focusing on retrieving and querying semantic relations between entities in archaeological literature. These experiments are still in an early stage of development, but they represent an important extension of existing approaches on grey literature in archaeology, which focus on identifying and classifying archaeologically relevant entities rather than relations between them. In the longer perspective they will thus have a profound impact on the way scientific literature is accessed in archaeology and other branches of Digital Humanities.

So far, our experiments have been conducted on English texts only. But this is only because of the availability of Natural Language Processing resources for this particular language, which makes it a promising candidate for initial experiments and for determining which technologies to choose for our specific task. With information extraction experiments successfully conducted on that basis, analogous processing pipelines for other languages can be created.

2 Use Case and Technological Background

In the scenario detailed in this paper, we imagine an archaeologist interested in objects that are described as having been *'found'* in the course of an excavation. We would like to emphasize that the example use case is different from the state of the art in Natural Language Processing as currently conducted on scientific publications and grey literature from archaeology which is represented by Named Entity Recognition and Entity Linking (Binding, Tudhope and May 2008, Byrne and Klein 2010). We go beyond detecting and classifying archeologically relevant terms in isolation, a task which we consider to be solvable by existing initiatives and their technologies. Instead,

we are interested in recovering semantic relations connecting such terms.

For reasons of space, we cannot provide an exhaustive introduction for the technologies described here, which fall under the broad scope of Natural Language Processing (NLP), Human Language Technologies (HLT) in its specific application to Digital Humanities (DH), as well as the Semantic Web (SW) resp. Linked Open Data (LOD)[1]. We see our activities within the more general scope of *Machine Reading* (Etzioni, Bank and Cafarella 2006) in its application to scientific publications and grey literature from the field of archaeology.

Following Barker (2007), Deep Machine Reading aims to provide a formal representation of a given text, say, a text book, as exhaustively as possible. It is related to concepts like traditional *Information Extraction* and *Text Mining*, but goes beyond these in that we aim to process not only information defined in pre-existing vocabularies or registries, but also evaluate *free text*. Unlike general-purpose *Open Information Extraction*, however, we formalize the output of our system in line with standards, technologies and logics developed in the context of the Semantic Web, thereby establishing not only machine-readable representation of the semantics of scientific publications, but also a representation with well-defined formal semantics, i.e., the Resource Description Framework RDF (W3C-RDF, 2014) and the Web Ontology Language OWL (W3C-OWL, 2012).

Using this representation, we aim to answer a query directly run against (the automatically extracted RDF representation of) a PDF document for a natural language question like *'What did they find?'*

3 Open Information Extraction: From PDF to RDF

For our first experiments we choose digital-born PDFs including selected publications of the Römisch-Germanische Kommision since 2004 (*Germania*, *Bericht der Römisch-Germanischen Kommission*), and FASTI Online (Fasti Online 2015).

Out of this pool of data we currently focus on Imperial Rome. The technology is, however, not specific to such data but may be applied to other strands of archaeology (and beyond). Below, we use Muccigrosso (2011) as an example text for the analysis.

3.1 Text Extraction

Extracting text from a digital publication designated for print is not a trivial issue. We extract text using PDF2XML (PDF2XML, 2015) and a set of tailored XSLT scripts which heuristically detect and classify textual content (titles, author, headlines and paragraphs), de-hyphenate line breaks and merge paragraphs across page breaks.

[1] For general introductions into these areas, we recommend Jurafsky (2008) for Natural Language Processing, Schreibman (2004) for Digital Humanities, Hitzler (2009) for Semantic Web technologies and Berners-Lee (2009) for Linked Data. Unless an explicit reference is given, the technical terms used below are used as defined in these works.

3.2 Natural Language Processing

The actual NLP pipeline takes the resulting text as its input and uses existing NLP tools for linguistic analysis of the text, including steps of sentence splitting, tokenization, part-of-speech tagging, lemmatization, syntactic parsing and named entity recognition. Particularly relevant for our example is the sub-task of *Semantic Role Labeling* (SRL, Palmer, Gildea, Kingsbury 2010): Semantic roles, or theta-roles, describe the semantic relationship between a predicate (say, a verb) and its arguments (say, subject and object), often formalized in terms of frame semantics. In this context, any particular frame consists of a number of semantically defined 'slots' for predicate and argument, but in addition, it is defined as being in a particular ontological relation with other frames, e.g., in terms of inheritance. Semantic Role Labeling, the task of automatically identifying predicates (both verbal and nominal), their arguments (e.g., prepositional phrases) and the semantic role between them, thus represents a major component in our approach.

The result of the NLP analysis for the example sentence marked blue in the figure above is shown below.

Here, the first column (coloured in an ascending red-green scale) is the number of the word in the sentence, the second column is the actual word, followed by lemma, parts of speech, named entities, shallow syntax (chunking) and a phrase structure parse. The 8th and 9th columns provide a dependency representation of this parse with links to the respective head of a given word (colours match the colours of the first column) and the respective dependency labels. Then, semantic role labelling follows with the list of predicates. The arguments of the first predicate (*basing*) are shown in the following column, those of the second in the one after, etc.

3.3 Target Format: Resource Description Framework (RDF)

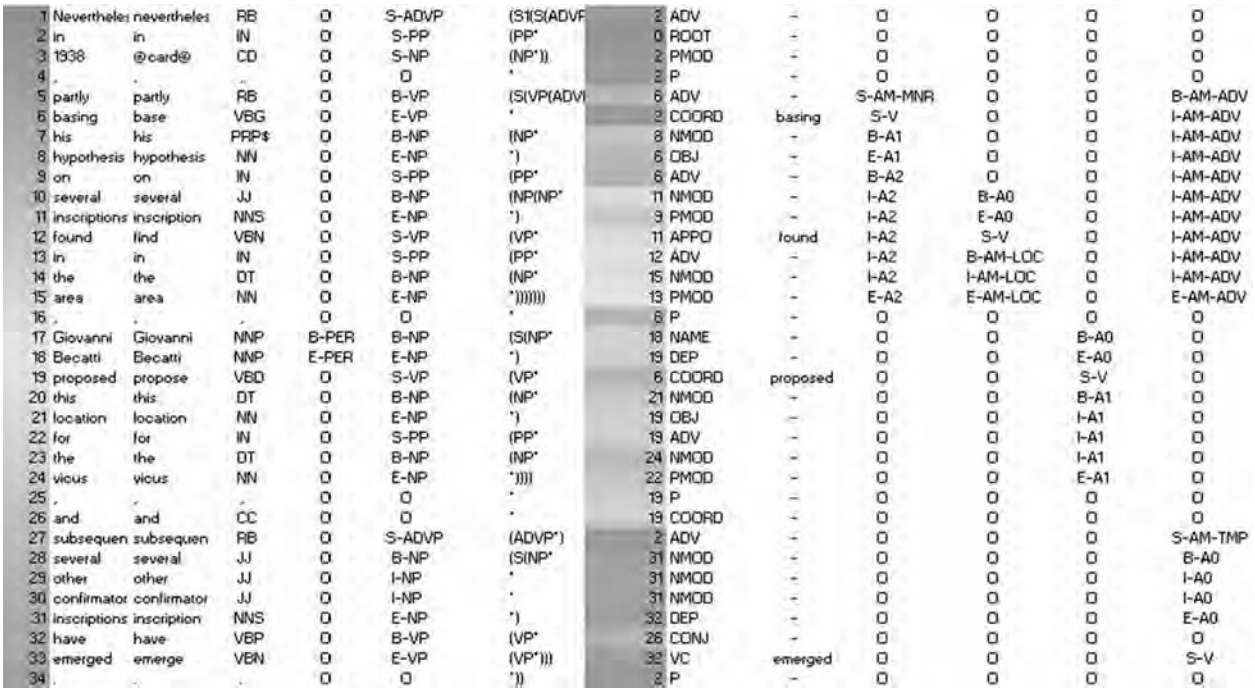
For representing the information to be extracted from the text, we adopt the Resource Description Framework (RDF, W3C-RDF 2014) as a modelling toolkit: The fundamental data structure of RDF is a *triple*, i.e., a pair of two *nodes* (RDF resources) connected by a labelled edge (RDF property/ predicate). Edges in this graph structure are *directed*, with the source node being conventionally referred to as the '*subject*' and the target node the '*object*' of a particular triple. Subject, object and predicate are by themselves RDF resources, and can be identified with a Uniform Resource Identifier (URI). As a result, any RDF resource is uniquely addressable in the web (of data), e.g., in the form of a HTTP link. An example triple conveying the information that *We (continue to) find a relatively large number of coins* (Muccigrosso 2011) may thus have the following form:

:we :find : coins.

Additional triples then may further describe*:coins*, etc., e.g., as having the string representation *'a relatively large number of coins*'.

:coins rdfs:label 'a relatively large number of coins'^^xsd:string.

If URIs are resolvable (i.e., if a HTTP link opened in a browser or crawler points to a resource than can be accessed via HTTP



1	Nevertheles	nevertheles	RB	O	S-ADVP	(S1(S(ADVP	2	ADV	-	O	O	O	O
2	in	in	IN	O	S-PP	(PP*	0	ROOT	-	O	O	O	O
3	1938	@card@	CD	O	S-NP	(NP*))	2	PMOD	-	O	O	O	O
4	,	,	,	O	O	*	2	P	-	O	O	O	O
5	partly	partly	RB	O	B-VP	(S(VP(ADVI	6	ADV	-	S-AM-MNR	O	O	B-AM-ADV
6	basing	base	VBG	O	E-VP	*	2	COORD	basing	S-V	O	O	I-AM-ADV
7	his	his	PRP$	O	B-NP	(NP*	8	NMOD	-	B-A1	O	O	I-AM-ADV
8	hypothesis	hypothesis	NN	O	E-NP	*)	6	OBJ	-	E-A1	O	O	I-AM-ADV
9	on	on	IN	O	S-PP	(PP*	6	ADV	-	B-A2	O	O	I-AM-ADV
10	several	several	JJ	O	B-NP	(NP(NP*	11	NMOD	-	I-A2	B-A0	O	I-AM-ADV
11	inscriptions	inscription	NNS	O	E-NP	*)	9	PMOD	-	I-A2	E-A0	O	I-AM-ADV
12	found	find	VBN	O	S-VP	(VP*	11	APPO	found	I-A2	S-V	O	I-AM-ADV
13	in	in	IN	O	S-PP	(PP*	12	ADV	-	I-A2	B-AM-LOC	O	I-AM-ADV
14	the	the	DT	O	B-NP	(NP*	15	NMOD	-	I-A2	I-AM-LOC	O	I-AM-ADV
15	area	area	NN	O	E-NP	*)))))))	13	PMOD	-	E-A2	E-AM-LOC	O	E-AM-ADV
16	,	,	,	O	O	*	6	P	-	O	O	O	O
17	Giovanni	Giovanni	NNP	B-PER	B-NP	(S(NP*	18	NAME	-	O	O	B-A0	O
18	Becatti	Becatti	NNP	E-PER	E-NP	*)	19	DEP	-	O	O	E-A0	O
19	proposed	propose	VBD	O	S-VP	(VP*	6	COORD	proposed	O	O	S-V	O
20	this	this	DT	O	B-NP	(NP*	21	NMOD	-	O	O	B-A1	O
21	location	location	NN	O	E-NP	*)	19	OBJ	-	O	O	I-A1	O
22	for	for	IN	O	S-PP	(PP*	19	ADV	-	O	O	I-A1	O
23	the	the	DT	O	B-NP	(NP*	24	NMOD	-	O	O	I-A1	O
24	vicus	vicus	NN	O	E-NP	*))))	22	PMOD	-	O	O	E-A1	O
25	,	,	,	O	O	*	19	P	-	O	O	O	O
26	and	and	CC	O	O	*	19	COORD	-	O	O	O	O
27	subsequen	subsequen	RB	O	S-ADVP	(ADVP*)	2	ADV	-	O	O	O	S-AM-TMP
28	several	several	JJ	O	B-NP	(S(NP*	31	NMOD	-	O	O	O	B-A0
29	other	other	JJ	O	I-NP	*	31	NMOD	-	O	O	O	I-A0
30	confirmator	confirmator	JJ	O	I-NP	*	31	NMOD	-	O	O	O	I-A0
31	inscriptions	inscription	NNS	O	E-NP	*)	32	DEP	-	O	O	O	E-A0
32	have	have	VBP	O	B-VP	(VP*	26	CONJ	-	O	O	O	O
33	emerged	emerge	VBN	O	E-VP	(VP*)))	32	VC	emerged	O	O	O	S-V
34	.	.	.	O	O	*))	2	P	-	O	O	O	O

FIG.1.

and that provides information about the resource), then data sets on different remote servers share identifiers for, e.g., terminology, and provide cross-links with each other, a concept conventionally known as 'Linked (Open) Data' (Berners-Lee, Bizer and Heath 2009). With the concept of *federation*, SPARQL 1.1 (W3C-SPARQL 2013) allows to query such links across distributed resources, so that the set of interlinked resources accessible in this way and available under an open license forms the 'Linked Open Data (LOD) cloud'. This is an interesting feature when it comes to combining different knowledge sources (Section 5).

We don't use the RDF-toolkit to represent the domain by itself, but the lexical semantics of the text. Aggregating multiple texts will, however, approximate a domain model in that we can extrapolate typical properties and their likelihood to connect (instances of) specific classes. Our presentation featured such a concept graph bootstrapped from analysing raw text, omitted here caused by the format of the paper. While such an automatically constructed domain model is from a quite different quality than a formal ontology, it can be used to infer additional information (Penas and Hovy 2010).

3.4 Triple Extraction

To construct RDF triples out of plain text, we ground most of our triple extraction on 'Semantic Role Labeling' a la PropBank (Palmer, Gildea and Kingsbury 2005). Unlike other representation formalisms for semantic roles, PropBank limits itself to a *minimal set* of semantic roles and aims for a high degree of genericity. For every verbal predicate, it distinguishes 6 classes of direct arguments, the most important being A0 (AGENT, prototypical subject), A1 (PATIENT, prototypical direct object), and A2 (THEME, prototypical indirect object), whereas the other classes are predicate-specific. In addition, several classes of oblique arguments are supported, including AM-NEG (negative modifier), AM-LOC (locative modifier) and AM-TMP (temporal modifier).

For every transitive verb, then, a triple is formed connecting its main arguments (A0 and A1) with a relation that carries the lemma of the verbal predicate, resulting in the *:proposes* relation in Figure 2. Other semantic roles are then connected to A0 and A1 arguments with relations composed of the generic relation identifier *:do* combined with a placeholder for the respective semantic role, e.g. *at* for AM-LOC and *during* for AM-TMP. Hence, we establish the relation *:do-at* between A0 arguments and locative modifiers, etc.

Figure 2 gives a full example analysis of the sentence '*Nevertheless in 1938, partly basing his hypothesis on several inscriptions found in the area, Giovanni Becatti proposed this location for the vicus , and subsequently several other confirmatory inscriptions have emerged*' in graphical form. (Note that here, rdfs:labels replace the actual URIs and cardinality properties have been omitted.)

This fragment captures roughly the following semantics:

Giovanni Becatti proposes 'the vicus'

someone (_:n5) bases 'his hypothesis' on inscriptions found by someone (_:n4) in 'the area'

'confirmatory inscriptions' emerged

An obvious limitation of this representation is that context-dependencies get not resolved. Of course, the blank node *_:n5* is to be resolved to Giovanni Becatti, who also occurs to possess 'his hypothesis'. But in addition, *the vicus* and *the area* need to be identified with areas or vici mentioned before, and the relational nature of the adjective *confirmatory* (which

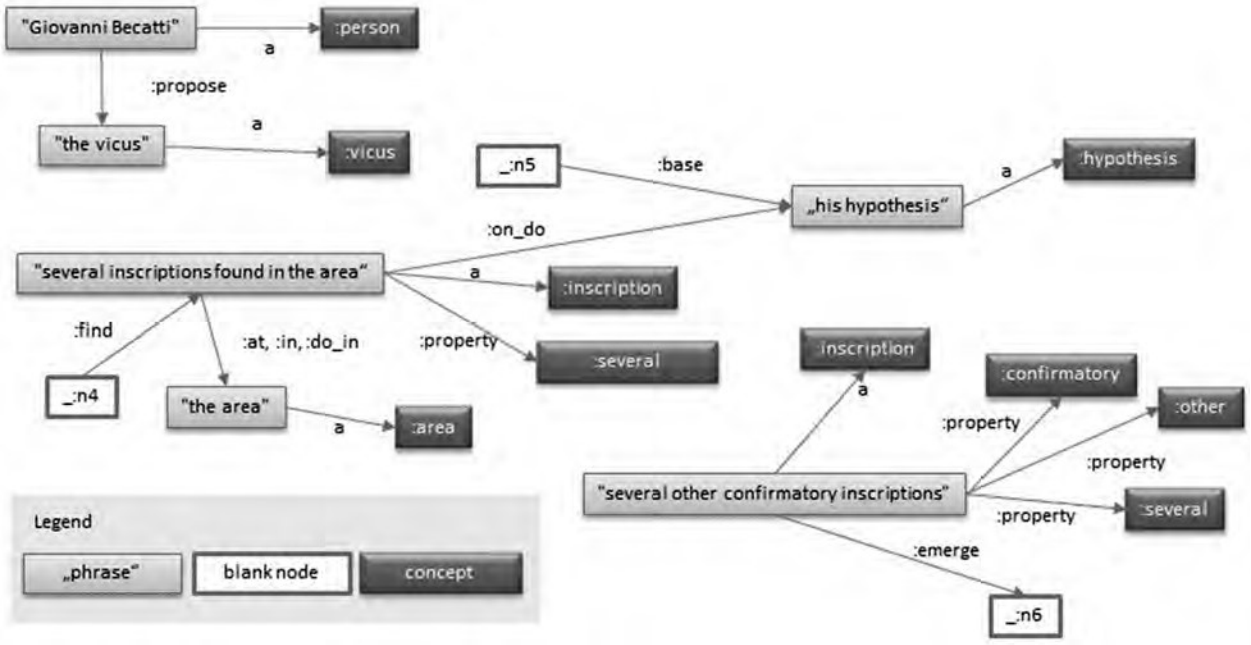


FIG. 2.

presupposes a hypothesis to be confirmed) is not recognized. Partially, these problems can be handled through anaphor resolution systems. At the moment, the development or domain adaptation of such a system is beyond the scope of our initial experiments, but future refinement of the extracted information will include the support for co-reference.

4 Querying RDF

The extracted data can be directly queried using the SPARQL (W3C-SPARQL 2013), the standard query language for RDF data. However, as this may be inconvenient to archaeologists, we also provide a simple (though limited) natural language query interface to the extracted data. The general idea is appealingly simple: Given a user query, run the NLP pipeline and triple extraction procedure above and convert the output into a SPARQL query by replacing object and subject URIs by variables.

As an example, the analysis of the query *What did they find?* is shown below, together with its SPARQL version.

The resulting query is just a minor, and fully automated modification of the triples generated from the NLP analysis: SELECT and WHERE statements are added as obligatory components of the query, the arguments of *:find* are transformed from URIs to variables (marked by *?*) and the string values of their labels are replaced by variables, as well. For the result, SELECT requires that only these label variables are returned.

This trivial transformation works already well, and it returns matches for phrases like *X found Y*, *Y has been found by X,* or *X, the finder of Y*. With this naive approach, however, it is not possible to query for optional arguments (unless every variable is set to optional by default), and hence, the result set is limited to instances of *:find* that come with an explicit A0 argument. To query for objects of *:find*, only one can, however, ask *What was found?*

So far, this system remains bound to lemma matches with the text. While *What did they find?* generates the same SPARQL query as *Who found what?*, the system is not capable yet to capture the generalization to *X discovered Y*.

5 Background Knowledge and Inference

To unleash the potential of semantic technologies, inference beyond plain lemma matches as described above is a key requirement. This involves augmenting the extracted data with semantic information both for general world knowledge and for domain-specific vocabulary. As we are interested in verbal predicates here, we describe linking with an existing resource for verbal semantics. Other lexical-conceptual resources, however, can be processed analogously[2].

Similar to concepts during Named Entity Recognition, resp. Entity Linking, properties can be linked, inferred and queried. As we do not preserve their property labels during triple extraction, though, we rely on *URI match* during linking, i.e., the use of identical predicates. VerbNet (Kipper *et al.* 2006) is an extension of the verbal lexicon which provides a taxonomy of verb senses, with leaves representing sets of (English) verbs, as it takes the syntactic realization of arguments into account, we used this resource to generalize over predicates.

However, VerbNet identifiers are partially abstract. Accordingly, we chose not to query for them directly, but to assume that all verbs associated with the same concept are (to a certain extent)

[2] In addition to the experiment described in the text, we created and experimented with two small domain vocabularies, i.e., a minimal Dutch-English-German SKOS vocabulary of archeological features and periods (29 concepts), and a German thesaurus of 4552 concepts for classical archeology covering general excavation terminology (265 concepts), Greek/Roman mythology, toponyms and ethnonyms (1430 concepts), artifacts (1192 concepts) and materials (242 concepts), architecture (457 concepts), as well as anthropology, botanics and zoology (561 concepts).

```
#What did they find ?
       data:n0_000_what        rdfs:label      "What"^^xsd:string.
       data:n0_002_they        :find           data:n0_000_what.
       data:n0_002_they        rdfs:label      "they"^^xsd:string.
```

```
SELECT DISTINCT ?n0_000_what_label ?n0_002_they_label
WHERE {
       ?n0_000_what            rdfs:label      ?n0_000_what_label .
       ?n0_002_they            :find           ?n0_000_what .
       ?n0_002_they            rdfs:label      ?n0_002_they_label .
}
```

TAB. 1.

```
PREFIX terms: <http://purl.org/acoli/open-ie/>
...
SELECT DISTINCT ?n0_000_what_label ?do0
WHERE {
       _:n2 :declare-29.4|:discover-84|:get-13.5.1  ?n0_000_what .      # superproperties of :find
       _:n2 ?do0 ?n0_000_what .                                        # which predicate is actually used ?
       FILTER regex(str(?do0), "^http://purl.org/acoli/open-ie/[^0-9]*$")# limit to terms predicates
       ?n0_000_what rdfs:label ?n0_000_what_label .
}
```

TAB. 2.

n0_000_what_label	do0
"no new inscriptions naming the site"	<http://purl.org/acoli/open-ie/discover>
"The second tomb discovered in 2010"	<http://purl.org/acoli/open-ie/discover>
"several classes of object which indicate the presence of fairly elaborate buildings at the site"	<http://purl.org/acoli/open-ie/find>
"the fragment found last year from the factory of Suriscus"	<http://purl.org/acoli/open-ie/find>
"a relatively large number of coins"	<http://purl.org/acoli/open-ie/find>
"a deposit of what we believe to be mortar , which may have come from the now missing upper ...	<http://purl.org/acoli/open-ie/find>
"several drainage trenches running EW across the site , cutting through ancient walls and other features"	<http://purl.org/acoli/open-ie/discover>
"Their location along the Via Flaminia just 3.5 km from our site makes plausible their use by the ...	<http://purl.org/acoli/open-ie/find>

FIG. 3.

semantically similar, and broadened the query to all sibling verbs of the actual verb queried. Sibling verbs, as defined here, are children of the immediate parent node(s) of the verb we queried for. An example query explicitly addressing the sibling concepts is shown below. According to VerbNet, the verb 'find' is found in the verbal senses *:declare-29.4*, *:discover-84* and *:get-13.5.1*. Using this generalization, however, we lose information about the actual verb used, so that we add an additional variable to the query, limit its values to URIs from the *terms* namespace in which our extracted properties and those of VerbNet reside and include it in the result.

If we allow the query generation engine to access the VerbNet hierarchy, the SPARQL example below can be generated for *What has been found?*

Now, this query not only retrieves results for *find*, but also for, e.g., *discover*:

Note that this query requires RDFS reasoning and thus require enabling the corresponding entailment regime in the database. As an alternative, direct querying in native RDF is possible by means of SPARQL 1.1 property paths: Assuming that the property *:find* is defined as an *rdfs:subPropertyOf :declare-29.4* etc. (in accordance with its lexeme information in VerbNet), the following query is equivalent to the one given above.

Here, we use the original predicate *:find* as an anchor, we go one step up to its VerbNet generalization(s) (defined with using *rdfs:subPropertyOf*) and one step down to its sibling concepts (using the inverse property path *^rdfs:subPropertyOf*). The relation between superproperties and properties created from the text (i.e., verbal lemmas) is drawn from VerbNet, and as here, only the lowest level in the VerbNet hierarchy is addressed, it is not necessary to limit the result to the *term:* namespace anymore. At this level, the only sub-properties must have come from the text, i.e., the *term:* namespace. The result of the query, however, remains the same, but with the inference (i.e., access to the VerbNet hierarchy) handled internally by the RDF data base rather than an explicit VerbNet lookup. Like the query in Sect. 4, this query can thus be *automatically* generated from the analysis of a natural language question.

Similar semantic generalizations are possible when concepts are addressed and a terminological resource with hierarchical structure is employed, e.g., WordNet-RDF (2015).

This experiment shows that semantically supported access to archaeological publications is possible and promising, and that

```
PREFIX : <http://purl.org/acoli/open-ie/>
PREFIX terms: <http://purl.org/acoli/open-ie/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?n0_000_what_label ?do0
WHERE {
        :find rdfs:subPropertyOf/^rdfs:subPropertyOf ?do0 .      # ?do0 is a sibling property of :find
        _:n2 ?do0 ?n0_000_what .                                 # the triple in the data
        ?n0_000_what rdfs:label ?n0_000_what_label .
}
```

Tab. 3.

it can be supported with existing resources for, e.g., the general semantics of verbs. Both observation are, however, merely subjective impressions at the moment and require a more in-depth evaluation within a concrete application scenario.

6 Conclusions and Prospects

Our experiments show that the application of state-of-the-art semantic technologies from both NLP and Semantic Web is both possible and potentially fruitful for developing innovative applications of methodological value to archaeologists as well as other fields of (Digital) Humanities that are at least in parts concerned with scanning and accessing existing collections of heterogeneous scientific text.

The NLP analysis and triple generation is implemented as described. On this basis, we conducted additional experiments using off-the-shelf technology. None of this is integrated into a toolkit tailored towards end users, but it requires a minimal level of technical background to be replicated. Our point is to show how easily these experiments could be conducted with minimal knowledge of SPARQL on the side of the user, and this is a basis for developing concrete tools.

Core functionalities such as basic query interfaces and subsumption inferences are already available or can be easily developed. Also, general lexical resources seem applicable to the domain of archaeology. Nevertheless, the development of domain vocabularies, or the development of bootstrapping domain vocabularies from the existing body of text is a desideratum of great importance

A fundamental problem here is that ontological resources for the archaeology which are (or are supposed to be) developed at or by larger initiatives (e.g., Ariadne) are rarely publicly available. The freely accessible ontologies we are aware of are highly domain-specific (e.g., http://nomisma.org for Roman numismatics, http://data.archaeologydataservice.ac.uk/page/ for datasets and publications) or provide only TBox information (http://www.heritagedata.org/crmeh/crmeh_current.rdf) for documenting excavations. We do not have an ontology of find-spots, named entities, archaeological features, cultures, etc. which could be used for this purpose in a sufficient way.

So far we did not tackle the problem of multilinguism. Ontology localization has been a major topic in the Semantic Web community, e.g., in the context of the OntoLex and Multilingual Semantic Web workshops or in the OntoLex W3C community group founded in 2012 (OntoLex 2015). A simple solution that maintains queriability (at the expense of a non-minimal and possibly incorrect representation) would be to use bilingual word lists to 'translate' concept and property names created from one foreign-language text to the target language (say, English), with all possible translations generated out. A more advanced solution would be context-aware statistical machine translation, an idea currently pursued in the above-mentioned community efforts.

Additionally, our NLP components need to be extended to other languages. Thinking about German and French, this is a relatively easy task for languages with such richly developed research landscapes in the field of NLP, for other European (and even worse, non-European) languages, however, the situation is more problematic. This already includes Dutch (for which we possess neither Semantic Role Labelling nor an anaphor resolution system), but for other European languages, the situation is even worse (META-NET 2015). A technical problem in this regard is that NLP technology has a traditional focus on English whose lack of morphology is particularly suitable for the development of statistical NLP tools. The development of elementary tools for morphology-rich languages (say, Slavic, Greek, Latin, Finnish, Hungarian, Turkish, or Arabic) is still an active area of research. For the immediate future, we thus focus on selected languages with substantial NLP support (English, German, possibly French). In addition, we aim to experiment with a mid-resourced language, Dutch in our example, to assess the potential and the efforts required to extend the coverage of languages beyond this immediate core group.

Another aspect is that additional NLP techniques need to be integrated. In particular, an anaphor resolution system to facilitate information aggregation across sentence boundaries. As anaphor resolution benefits from rich semantic information, we aim to adapt an existing anaphor resolution system to take domain-specific information into account. While such efforts are beyond the scope of the pilot studies described here, they should be a major component of any more dedicated project.

To conclude, we developed and described prototypical core components of a system for machine reading scientific texts, illustrated here with a novel application in the field of archaeology, and sketched their application to search for relations in this body of text. Using formal background knowledge (VerbNet in the example), we were able to answer a natural language query in a way that not only literal matches for the verb *find* could be retrieved, but also matches from related verbal concepts like *discover*. Although no archaeology-specific resources were employed in this example, we have demonstrated the principal applicability of our technologies to this domain. At the same time, major technical problems (coreference, performance optimization, modal and contextual information, multilingualism) remain to be addressed, these are to be addressed within the scope of a dedicated research project that combines an original research problem from archaeology with this kind of technology to demonstrate its benefits and potential and to facilitate its adaptation by the scientific community. At the same time, any such project should provide expert knowledge from archaeologists for the automatically assisted creation, curation and extension of terminology resources.

Bibliography

Barker, K., Agashe, B. *et al.* 2007. Learning by reading: A prototype system, performance baseline and lessons learned. In *Proceedings of the Twenty-Second National Conference on Artificial Intelligence (AAAI 2007, Vancouver, British Columbia)*: 280-6.

Binding, C., May K., Tudhope D. 2008. Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. *In Proceedings of European Conference on Digital Libraries* (ECDL08), no. 5173 (Aarhus, Denmark, Springer): 280–90.

Berners-Lee, T., Bizer C., and Heath, T. 2009. *Linked data – The story so far*. International Journal on Semantic Web and Information Systems 5, 3: 1-22.

Byrne, K. and Klein, E. 2010. Automatic Extraction of Archaeological Events from Text, in Frischer, B., Webb Craford, J., and Koller, D. (eds.), *Making History Interactive. Computer Applications and Quantitative Methods in Archaeology* (CAA). Proceedings of the 37th International Conference, Williamsburg, Virginia, United States of America, March 22-26 (BAR International Series S 2079): 48-56. Archaeopress, Oxford.

Etzioni, O., Banko, M., and Cafarella, M. 2006. Machine Reading. In *Proceedings of the 21st Conference on Artificial Intelligence (AAAI-2006, Boston)*: 1517-9.

Fasti Online 2015. A database of archaeological excavations since the year 2000 [online] http://www.fastionline.org/index.php [Accessed: 29th June 2015].

Hitzler, P., Krötzsch M. and Rudolph S. 2009. Foundations of Semantic Web technologies. Boca Raton, CRC Press.

Jurafsky, D. and Martin, J. 2008. Speech and language processing. Upper Saddle River, Prentice Hall.

Kipper, K., Korhonen A., Ryant A. and Palmer, M. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC-2006*, Genoa.

Muccigrosso, J. D. 2011. The 2010 Excavation Season at the Site of the Vicus ad Martis Tudertium *(PG)*, *The Journal of Fasti Online, Associazione Internazionale di Archeologia Classica*, http://www.fastionline.org/docs/FOLDER-it-2011-227.pdf [Accessed: 29 june 2015].

META-NET 2015. A Network of Excellence forging the Multilingual Europe Technology Alliance [online] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison [Accessed: 29 june 2015].

ONTOLEX 2015. Ontology-Lexica community group [online] https://www.w3.org/community/ontolex/ [Accessed: 29 june 2015].

Palmer, M., Gildea, D. and Kingsbury, P. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1): 71-106.

Penas, A. and Hovy, E. 2010. Semantic Enrichment of Text with Background Knowledge
http://www.aclweb.org/anthology/W10-0903.pdf [Accessed: 15 november 2015].

Palmer, M., Gildea D. and Xue, N. (2010) Semantic role labelling. Synthesis Lectures on Human Language Technologies 3(1): 1-103. San Rafael, Morgan and Claypool.

PDF2XML 2015. Pdf2xml converter based on Xpdf library [online] http://sourceforge.net/projects/pdf2xml [Accessed: 29 june 2015].

Schreibman, S., Siemens, R. Unsworth, J. 2004. A Companion to Digital Humanities. Oxford, Blackwell.

W3C-OWL 2012. OWL 2 Web Ontology Language (Second Edition), W3C Recommendation 11 December 2012 [online] http://www.w3.org/TR/owl2-overview [Accessed: 29 june 2015].

W3C-RDF 2014. RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation 25 February 2014 http://www.w3.org/TR/rdf11-concepts [Accessed: 29 june 2015].

W3C-SPARQL 2013. SPARQL 1.1 Overview, W3C Recommendation 21 March 2013 http://www.w3.org/TR/sparql11-overview/ [Accessed: 29 june 2015].

WORDNET-RDF 2015. WordNet RDF http://wordnet-rdf.princeton.edu/ [Accessed: 29 june 2015].